DOCUMENT RESUME

ED 429 999 TM 029 709

AUTHOR	Wang, Wen-Chung
TITLE	Decomposition of DIF Effects for Polytomous Items.
PUB DATE	1998-04-00
NOTE	26p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Foreign Countries; Identification; *Item Bias; Models; Simulation
IDENTIFIERS	*Decomposition Analysis (Statistics); *Polytomous Items

ABSTRACT

Conventional two-group differential item functioning (DIF) analysis for dichotomous items is extended to factorial DIF analysis for polytomous items where multiple grouping factors with multiple groups in each are jointly analyzed. By adopting the formulation of general linear models, item parameters across all possible groups are treated as a dependent variable and the factors as independent variables. These item parameters are then reparameterized as a set of grand item parameters and sets of DIF parameters representing main and interaction effects of the factors on the items. Results of simulation studies show that the parameters of the proposed modeling could be satisfactorily recovered. A real data set of 10 polytomous items with 1,924 subjects was analyzed. Applications and implications of the proposed modeling are addressed. (Contains 3 tables, 4 figures, and 18 references.) (Author)

Decomposition of DIF Effects for Polytomous Items

WEN-CHUNG WANG

National Chung Cheng University

Correspondence:
Department of Psychology
National Chung Cheng University
Chia-Yi 621, Taiwan
Telephone: 886-5-2720411 ext. 6430
Fax: 886-5-2720857
E-mail: psywcw@ccunix.ccu.edu.tw

Paper presented at the annual meeting of the American Educational Research Association, 1998.

Abstract

Conventional two-group DIF analysis for dichotomous items is extended to factorial DIF analysis for polytomous items where multiple grouping factors with multiple groups in each are jointly analyzed. By adopting the formulation of general linear models, item parameters across all possible groups are treated as a dependent variable and the factors as independent variables. These item parameters are then reparameterized as a set of grand item parameters and sets of DIF parameters representing main and interaction effects of the factors on the items. Results of simulation studies show that the parameters of the proposed modeling could be satisfactorily recovered. A real data set of 10 polytomous items and 1924 subjects was analyzed. Applications and implications of the proposed modeling are addressed.

Keywords: differential item functioning, polytomous item, Rasch model, partial credit model, general linear models, analysis of variance.

Procedures for detecting differential item functioning (DIF) for dichotomous items have been thoroughly investigated (Holland & Wainer, 1993). Recently, educational reform efforts have led to an increase use of polytomous items. Various procedures for the assessment of differential item functioning for polytomous items have also been proposed (Chang, Mazzeo, & Roussos, 1996; Dorans & Schmitt, 1993; Muraki, 1993; Rogers & Swaminathan, 1993; Welch & Hoover, 1993; Wilson, Spray, & Miller, 1993; Zwick, Donoghue, & Grima, 1993). Potenza and Dorans (1995) proposed a two-dimensional framework for classifying these approaches. On one dimension, an observed score or an estimate of a latent trait is used as a matching variable. On the other dimension, either a parametric approach or a nonparametric approach is used.

The latent-trait/parameteric approach is usually based on item response theory. For example, Lord (1980) pointed out that item characteristic curves are ideally suited to defining DIF. Since item parameters as well as person parameters determine the curves, the detection of DIF could be made by comparing item parameters for a focal group and a reference group. Take the Rasch model (Rasch, 1960) as an example. It suggests:

$$\log\left(\frac{p_{i1}}{p_{i0}}\right) = \theta_n - \delta_i, \qquad (1)$$

where p_{i0} and p_{i1} denote the probabilities of an incorrect answer (scoring 0) and a correct answer (scoring 1) to item i, respectively; θ_n denotes the ability of person n, and δ_i denotes the difficulty of item i. We could calibrate the item difficulties separately for each group. Then, the difference in item difficulties for two groups can be tested as follows:

$$Z_i = \frac{\hat{\delta}_{i1} - \hat{\delta}_{i2}}{\sqrt{Var(\hat{\delta}_{i1}) + Var(\hat{\delta}_{i2})}}, \qquad (2)$$

where $\hat{\delta}_{i1}$ and $\hat{\delta}_{i2}$ are maximum likelihood estimates of item i's difficulty for groups 1 and 2, respectively; $Var(\hat{\delta}_{i1})$ and $Var(\hat{\delta}_{i2})$ are their estimated error variances, respectively. Z_i follows approximately the standard normal distribution.

Since the estimation of the standard errors is usually imprecise, Thissen, Steinberg and Wainer (1988) adopted a marginal maximum likelihood (MML) estimation with the EM algorithm (Bock & Aitkin, 1981) to investigate DIF. A full model where different groups have different item difficulties was formed. In the framework of the Rasch model, the full model looks like:

$$\log\left(\frac{p_{i1}}{p_{i0}}\right)_k = \theta_n - \delta_{ik}, \tag{3}$$

where subscript k denotes group membership; δ_{ik} denotes the difficulty of item i for group k; and the others are defined as above. A reduced model where different groups yield the same item difficulties was also formed. The usual likelihood ratio test was then used to test the difference between these two nested models.

For polytomous items, the above two approaches can be directly extended. With the partial credit model (Masters, 1982), we may calibrate data of each group consecutively and then compare the ratio of the differences of step difficulties for two groups over its standard error to the standard normal distribution, as Equation (2). Or we may form a reduced model where the item step parameters are identical across groups and a full model where the item step parameters are different for different groups. Specifically, in the reduced model, we analyze the whole data set with the partial credit model:

$$\log\left(\frac{p_{ij}}{p_{ij-1}}\right) = \theta_n - \delta_{ij} \tag{4}$$

where p_{ij} denotes the probability of scoring j in item i; p_{ij-1} denotes the probability of scoring j - 1 in item i; θ_n denotes the ability of person n; δ_{ij} denotes the jth step difficulty of item i. In the full model, Equation (4) is extended to:

$$\log\left(\frac{p_{ij}}{p_{ij-1}}\right)_k = \theta_n - \delta_{ijk} \tag{5}$$

where δ_{ijk} denotes the jth step difficulty of item i for group k; the others are defined as above. The likelihood ratio test can then be applied to test the difference of the two models.

Consider there is more then one grouping factor (e.g., gender and ethnicity). We may treat all possible combinations of groups as levels of a new unified factor and apply the above DIF detection techniques. This approach has the disadvantage that the original grouping factors are invisible and the definition of the new unified factor is vague. Hu and Dorans (1989) found that deleting items for DIF can have unintended consequences for the groups that were not the focus of analysis. This finding leads to a marginal DIF analysis that the Educational Testing Service does. If there is more than one grouping factor such as gender and ethnic groups, instead of crossing one group factor with another to study DIF, they look at the margins. However, this marginal DIF analysis ignores potential interactions between these two factors. We need a procedure for DIF detection that not only reserves original grouping factors but also investigates interactions among factors.

The purpose of this study is to propose an approach that meets this demand. More specifically, the formulation of general linear models is adopted where item parameters are treated as a dependent variable and grouping factors as independent variables. Item parameters for all possible groups are reparameterized as a grand item parameter, sets of parameters representing main effects of the factors, and sets of parameters representing interaction effects among the factors. If these parameters for the main or the interaction effects are statistically different from zero, DIF is found. Moreover, these parameters depict the sizes of effects of the factors on the items. Thus, they are called DIF parameters. In the following sections, the formal parameterization is formulated. Item response models and computer software needed for this parameterization are introduced. Results of simulation studies are shown to draw that the grand item parameters and the DIF parameters can be satisfactorily recovered. Analysis of a real data set is also be provided. Finally, applications and implications of the study are discussed.

Parameterization of Item Parameters

Conventional two-group DIF analysis is analogous to the *t*-test for two independent means. As the *t*-test can be extended to simple and factorial analysis of variance (ANOVA) or general linear models (GLM), two-group DIF analysis can be extended to multiple-group analysis or multiple-factor/multiple-group DIF analysis. To begin with, let there be one factor with K groups, indexed $k = 1, \ldots, K$. Applying the partial credit model, we can estimate a set of step difficulties for each group separately, as shown in Equation (5). These item step difficulties δ_{ijk} can be reparameterized as:

$$\delta_{ijk} = \delta_{ij\cdot} + \alpha_{ijk}, \tag{6}$$

subject to the usual restrictions in GLM:

$$\sum_k \alpha_{ijk} = 0 .$$

With this formulation, $\delta_{ij\cdot}$ is in fact the average of the step parameters across K groups and thus represents the grand step difficulty of the jth step in item i; α_{ijk} is the deviation to the average and represents the effect of group k on the jth step difficulty of item i. It is a DIF parameter. Combining Equations (5) and (6) leads to:

$$\log\left(\frac{p_{ij}}{p_{ij-1}}\right)_k = \theta_n - (\delta_{ij} + \alpha_{ijk}).$$

If anyone of α_{ijk} for item i is significantly different from zero, the item exhibits DIF. To test this hypothesis, we can either compare the ratio of α_{ijk} over its estimated standard error to the standard normal distribution, or apply the likelihood ratio test to compare a full model with DIF parameters and a reduced model without DIF parameters.

Equation (6) and its accompanying restriction are analogous to simple ANOVA. As simple ANOVA can be extended to factorial ANOVA, Equation (6) can also be done. Consider there are two factors: Factor A with K levels, indexed $k = 1, \ldots, K$, and Factor B with L levels, indexed $l = 1, \ldots, L$. Altogether there would be $K \times L$ groups. Applying the partial credit model for each group consecutively, we could estimate the item step difficulties for as follows:

$$\log\left(\frac{p_{ij}}{p_{ij-1}}\right)_{kl} = \theta_n - \delta_{ijkl} \tag{7}$$

where subscript kl denotes group membership; δ_{ijkl} is the jth step difficulty of item i for group kl; the others are defined as above. Like the reparameterization in Equation (6), these item parameters can be reparameterized as:

$$\delta_{ijkl} = \delta_{ij\cdot} + \alpha_{ijk} + \beta_{ijl} + \alpha\beta_{ijkl}, \tag{8}$$

subject to the restrictions:

$$\begin{aligned} &\textstyle\sum_k \alpha_{ijk} = 0, \\ &\textstyle\sum_l \beta_{ijl} = 0, \\ &\textstyle\sum_k \alpha\beta_{ijkl} = \sum_l \alpha\beta_{ijkl} = 0. \end{aligned} \tag{9}$$

Combining Equations (7) and (8) leads to:

$$\log\left(\frac{p_{ij}}{p_{ij-1}}\right)_{kl} = \theta_n - (\delta_{ij\cdot} + \alpha_{ijk} + \beta_{ijl} + \alpha\beta_{ijkl}).$$

Consequently, $\delta_{ij\cdot}$ can be viewed as the grand *j*th step difficulty of item *i*, α_{ijk} as the main effect of Factor A_k, β_{ijl} as the main effect of Factor B_l, and $\alpha\beta_{ijkl}$ as the interaction effect of Factor A_k and Factor B_l, on the *j*th step difficulty of item *i*. α_{ijk}, β_{ijl}, and $\alpha\beta_{ijkl}$ are all DIF parameters. Equations (7), (8), and (9) can be directly generalized to more than two factors.

Estimation

The proposed procedure belongs to the Rasch family. Several existing Rasch models and their accompanying software can be used. The linear partial credit model (Fischer & Ponocny, 1994) with its accompanying software *LPCM* (Fischer & Ponocny, 1998) is an option. *LPCM* uses a conditional maximum likelihood estimation where no assumptions of person and item populations are needed. The software *ConQuest* (Wu, Adams, & Wilson, 1998) is another option. It was developed for the multidimensional random coefficients multinomial logit model (MRCML, Adams, Wilson, & Wang, 1997; Wang, Wilson, & Adams, 1997). MRCML is characterized by a scoring matrix and a design matrix. By manipulating the two matrices, the proposed factorial procedure can be implemented. *ConQuest* uses a marginal maximum likelihood estimation with the EM algorithm. A normal distribution is assumed (but not necessarily) for the person population. In the case of normal distribution, a mean and a variance for the person distribution and item parameters are jointly estimated. *ConQuest* is used in this study because it is user-friendlier for the proposed procedure.

With the MML estimation, only a grand population is assumed in the person facet. However, the groups analyzed may have quite different proficiency levels, that is, some groups may be more proficient than the others. Therefore, we have to parameterize the differences of group means in the item facet. If not, we are assuming all the groups come from the same population, which is very unlikely in practice. As in Equation (8), we may parameterize the means across groups as follows:

$$\mu_{kl} = \mu + \alpha_k^{'} + \beta_l^{'} + (\alpha\beta)_{kl}^{'}, \tag{10}$$

subject to the restrictions:

$$\sum_k \alpha_k^{'} = 0,$$

$$\sum_l \beta_l^{'} = 0,$$

$$\sum_k (\alpha\beta)_{kl}^{'} = \sum_l (\alpha\beta)_{kl}^{'} = 0.$$

With this formulation, μ_{kl} stands for the mean for group kl; μ stands for the grand mean; $\alpha_k^{'}$ stands for the deviation of the mean of Factor A_k to the grand mean; $\beta_l^{'}$ stands for the deviation of the mean of Factor B_l to the grand mean; $(\alpha\beta)_{kl}^{'}$ stands for the interaction of Factors A_k and B_l. The grand mean parameter μ as well as a common variance are modeled in the person facet. The other parameters, including the mean-deviation parameters ($\alpha_k^{'}$, $\beta_l^{'}$, and $(\alpha\beta)_{kl}^{'}$), the grand item parameters ($\delta_{ij\cdot}$), and the DIF parameters (α_{ijk}, β_{ijl}, and $\alpha\beta_{ijkl}$), are modeled in the item facet. With *ConQuest*, all the parameters are simultaneously estimated.

Simulation Studies

The design and the generating values of the simulation studies are based on the results of the following real data analyses. Two-way factorial design was adopted with two levels in each, which leads to four groups. The sample sizes of these four groups are 471, 476, 537, and 440. There are ten 3-point polytomous items. Two conditions were conducted: One is a full model with all possible DIF parameters estimated; the other is a reduced model with parts of the DIF parameters. One hundred replications were made under each condition.

Under the full model condition, altogether 81 parameters were estimated, including two person distribution parameters (a grand mean and a common variance), three mean-deviation parameters, 19 grand step difficulty parameters, 19 DIF step parameters for the main effects of Factor A_1, 19 DIF step parameters for the main effects of Factor B_1, and 19 DIF step parameters for the interaction effects of Factors A_1 and B_1. Table 1 summarizes the results of 100 replications: generating values, bias values (mean of hundred replications minus generating value), asymptotic standard errors, Z statistics (bias value divided by standard error), and root mean square errors (*RMSE*). According to the Z statistics, no parameters are significantly biased at the .05 level. All the parameters were recovered very well, with the bias values between -.021 and .015. Figure 1 depicts the relationship between the generating values and the bias values of the parameters. No systematic patterns are found.

Table 1. Generating values, bias values, asymptotic standard errors, Z statistics, and RMSE of various parameters in the full model

	Gen. Value	Bias	SE	Z	*RMSE*
Grand mean	.37	.0006	.0281	.022	.0281
Common variance	.77	.0081	.0350	.2313	.0359
Mean-deviation					
1	-.04	-.0017	.0266	.0637	.0266
2	.18	-.0007	.0236	.0299	.0236
3	-.02	.0003	.0267	-.0128	.0267
Grand step difficulty					
1_1*	-.32	.0031	.0582	.0524	.0583
1_2	1.22	.0121	.0613	.1982	.0625
2_1	-2.36	.0010	.1419	.0073	.1419
2_2	-1.10	-.0007	.0535	-.0134	.0535
3_1	-1.94	-.0012	.0783	-.0155	.0783
3_2	1.24	-.0006	.0573	-.0101	.0573
4_1	-1.27	-.0198	.0742	-.2664	.0768
4_2	.97	.0071	.0584	.1211	.0589
5_1	-.77	-.0043	.0648	-.0671	.0649
5_2	1.38	.0019	.0598	.0320	.0599
6_1	.30	.0006	.0477	.0128	.0477
6_2	3.94	-.0129	.1428	-.0905	.1434
7_1	-1.55	-.0095	.0834	-.1144	.0840
7_2	1.09	.0087	.0471	.1859	.0479
8_1	-1.79	-.0212	.0791	-.2674	.0819
8_2	.91	-.0023	.0533	-.0435	.0533
9_1	-.74	.0053	.0574	.0922	.0576
9_2	1.89	.0115	.0668	.1727	.0678
10_1	-1.86	-.0079	.0910	-.0871	.0913
Main effects of factor A_1(Males)					
1_1	-.07	.0041	.0604	.0681	.0606
1_2	-.14	.0003	.0681	.0040	.0681
2_1	.49	.005	.1404	.0354	.1405
2_2	.34	.0145	.0583	.2492	.0601
3_1	-.19	.0127	.0823	.1550	.0832
3_2	-.15	-.0027	.0591	-.0463	.0591
4_1	-.03	-.0051	.0805	-.0637	.0806
4_2	-.14	-.0100	.0471	-.2129	.0481
5_1	.02	-.0036	.0653	-.0554	.0654
5_2	-.10	-.0023	.0592	-.0385	.0593
6_1	.00	.0071	.0537	.1320	.0541
6_2	.05	-.0129	.1636	-.0788	.1641
7_1	.00	.0054	.0812	.0663	.0814
7_2	-.13	.0037	.0548	.0682	.0549
8_1	.02	-.0073	.0829	-.0876	.0832
8_2	.10	.0034	.0501	.0676	.0503
9_1	.11	-.0081	.0630	-.1287	.0635
9_2	.07	-.0107	.0674	-.1580	.0683
10_1	-.12	.0054	.0869	.0620	.0870
Main effects of factor B_1(Young)					
1_1	-.18	.0043	.0543	.0789	.0544
1_2	-.15	-.0055	.0593	-.0930	.0596
2_1	-.27	.0075	.1300	.0576	.1302
2_2	.26	.0079	.0538	.1466	.0543
3_1	.03	.0050	.0804	.0624	.0805
3_2	.18	-.0032	.0512	-.0631	.0513
4_1	.15	-.0033	.0683	-.0489	.0684
4_2	.31	-.0001	.0517	-.0016	.0517
5_1	.10	-.0005	.0644	-.0070	.0644
5_2	.28	.0014	.0568	.0246	.0568
6_1	-.02	-.0079	.0481	-.1642	.0487
6_2	-.07	-.0071	.1420	-.0496	.1422
7_1	.08	.0088	.0709	.1235	.0714
7_2	-.12	-.0053	.0559	-.0956	.0561
8_1	-.05	.0024	.0832	.0284	.0832
8_2	-.16	.0004	.0610	.0068	.0610
9_1	-.17	-.0006	.0583	-.0104	.0583
9_2	-.10	.0068	.0687	.0987	.069
10_1	-.03	-.0073	.1006	-.0726	.1009
Interaction effect of factors A_1 (Males) and B_1(Young)					
1_1	.07	.0004	.0578	.0063	.0578
1_2	.16	-.0005	.0703	-.0075	.0703
2_1	.14	-.0037	.1332	-.0277	.1332
2_2	.18	-.0013	.0649	-.0199	.0649
3_1	-.08	.0038	.0906	.0422	.0907
3_2	-.05	.0035	.0564	.0613	.0565
4_1	-.09	-.0156	.0576	-.2701	.0597
4_2	.04	-.0078	.0487	-.1597	.0493
5_1	-.12	-.0031	.0546	-.0569	.0547
5_2	.05	-.0009	.0587	-.0161	.0587
6_1	-.08	.0015	.0466	.0322	.0466
6_2	.15	-.0001	.1547	-.0005	.1547
7_1	-.15	-.0086	.0730	-.1178	.0735
7_2	.13	.0079	.0484	.1631	.0490
8_1	-.22	.0035	.0859	.0405	.0860
8_2	.00	.0123	.0512	.2400	.0526
9_1	-.03	.0015	.0592	.0251	.0592
9_2	-.13	.0045	.0657	.0681	.0658
10_1	-.07	.0042	.0829	.0513	.0830

* The first character denotes item number, and the second denotes step number. For example, 1_2 denotes the second step of item 1. This notation applies to other tables.

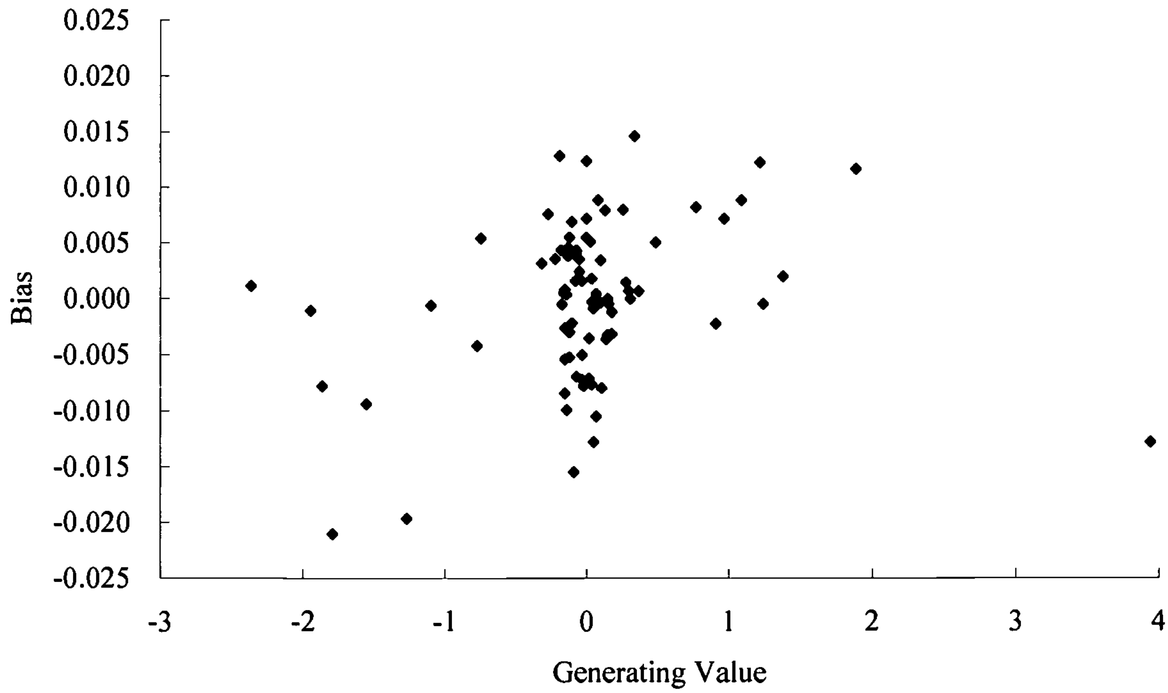

Figure 1. Parameter recovery under the full model condition

Under the reduced model condition, besides the above person distribution parameters, the mean-deviation parameters, and the grand step difficulty parameters, only 22 DIF step parameters were estimated, including seven DIF step parameters for the main effects of Factor A_1, ten DIF step parameters for the main effects of Factor B_1, and five DIF step parameters for the interaction effects of Factors A_1 and B_1. Results of 100 replications are summarized in Table 2. No parameters are significantly biased. All the parameters were recovered very well, with the bias values between -.045 and .043. Figure 2 displays the relationship between the generating values and the bias values. Again, no systematic patterns are found. In sum, under both conditions, all the parameters were recovered very well. This implies that the proposed modeling is not only theoretically preferable but also applicable.

Table 2. Generating values, bias values, asymptotic standard errors, Z statistics, and *RMSE* of various parameters in the reduced model

	Gen. Value	Bias	SE	Z	*RMSE*		Gen. Value	Bias	SE	Z	*RMSE*
Grand mean	.37	.0018	.0227	.0788	.0228	Main effects of factor A_1(Males)					
Common variance	.77	-.0003	.0405	-.0075	.0405	1_2	-.16	-.0095	.0675	-.1413	.0682
Mean-deviation						2_1	.43*	.0430	.1546	.2781	.1605
1	-.04	.0023	.0242	-.0942	.0243	2_2	.34*	-.0108	.0615	-.176	.0625
2	.18	.0035	.0324	-.1087	.0326	3_1	-.18	-.0010	.0865	-.0115	.0865
3	-.02	.0013	.0242	-.0544	.0242	3_2	-.15	-.0013	.0545	-.0237	.0545
Grand step Difficulty						4_2	-.15	-.0003	.0625	-.0053	.0625
1_1	-.32	.0088	.0572	.1532	.0579	7_2	-.13	-.0049	.0507	-.0969	.0510
1_2	1.22	.0043	.0609	.0699	.0610	Main effects of factor B_1(Young)					
2_1	-2.28	-.0363	.1445	-.2511	.1490	1_1	-.15	.0057	.0658	.0872	.0661
2_2	-1.10	.0039	.0603	.0647	.0604	1_2	-.11	-.0018	.0657	-.0272	.0657
3_1	-1.94	-.0032	.0834	-.0387	.0835	2_2	.27*	.0051	.0618	.0820	.0620
3_2	1.24	.0010	.0540	.0179	.0540	3_2	.22	.0064	.0596	.1066	.0599
4_1	-1.26	.0105	.0675	.1561	.0683	4_1	.17	-.0043	.0743	-.0582	.0745
4_2	.96	.0061	.0490	.1242	.0494	4_2	.34*	.0031	.0695	.0447	.0695
5_1	-.76	.0066	.0635	.1040	.0639	5_2	.35*	.0008	.0617	.0133	.0617
5_2	1.37	-.0023	.0557	-.0415	.0557	7_2	-.12	-.0451	.0585	-.7709	.0739
6_1	.31	-.0079	.0551	-.1435	.0556	8_2	-.13	-.0038	.0629	-.0601	.0630
6_2	3.9	.0068	.1264	.0535	.1266	9_1	-.16	.0007	.0673	.0111	.0673
7_1	-1.57	-.0074	.0705	-.1046	.0709	Interaction effect of factors A_1 (Males)and B_1 (Young)					
7_2	1.09	.0020	.0562	.0353	.0562	1_2	.20	.0067	.0609	.1104	.0612
8_1	-1.78	-.0044	.0771	-.0566	.0772	2_2	.20	-.0075	.0562	-.1340	.0567
8_2	.91	.0081	.0549	.1484	.0555	7_1	-.13	-.0052	.0762	-.0686	.0764
9_1	-.73	-.0061	.0549	-.1106	.0552	7_2	.14	-.0039	.0564	-.0688	.0566
9_2	1.88	.0152	.0721	.2101	.0737	8_1	-.21	.0049	.0825	.0595	.0827
10_1	-1.86	-.0075	.0825	-.0909	.0829						

* DIF effect is substantial according to Draba's recommendation

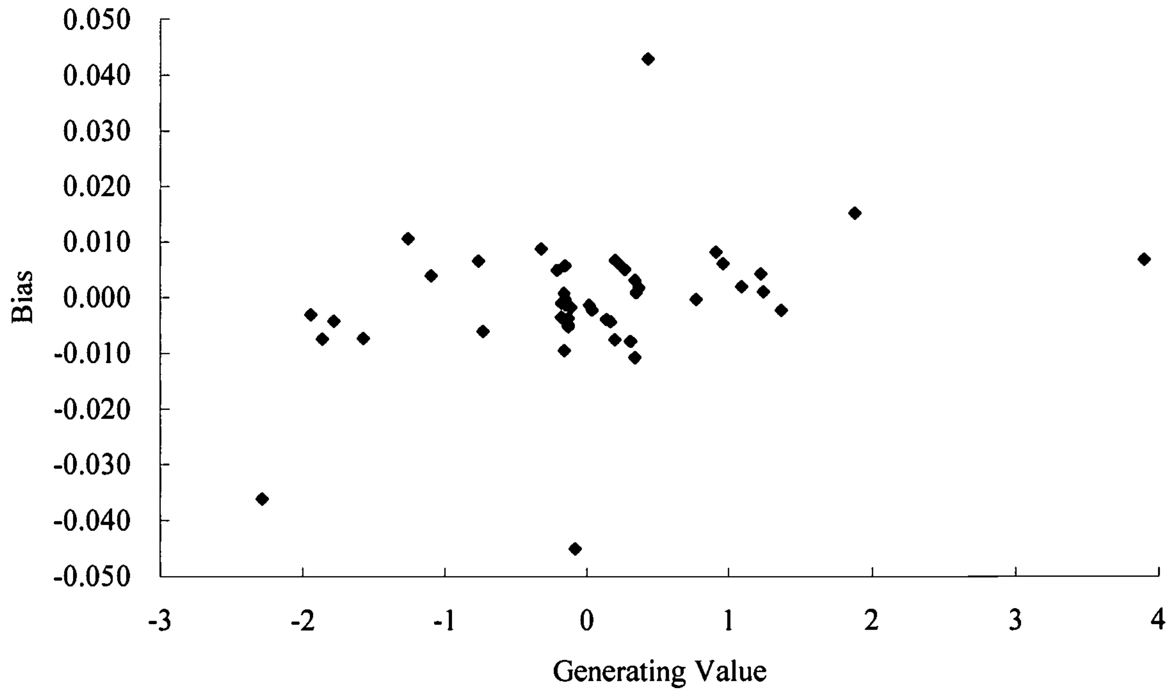

Figure 2. Parameter recovery under the reduced model condition

Real Data Analyses

The real data were collected by the research project "The social change in Taiwan, 1996." Ten 5-point Likert items (strongly agree, agree, undecided, disagree, strongly disagree) from an inventory of family values were analyzed. Subjects are 1924 adults in Taiwan. Because the sample sizes are not large, the categories "undecided", "disagree", and "strongly disagree" were combined into a new category "not agree". The categories "strongly agree", "agree", and "not agree" were scored 0, 1, and 2, respectively. Therefore, high scores indicate low values on family (i.e., more modern or liberal). Two factors are studied: Gender (factor A) and Age (factor B). Both factors have two levels: male (A_1) vs. female (A_2), young (B_1) vs. old (B_2). There are 471 young males, 476 young females, 537 old males, and 440 old females.

The partial credit model was first applied to the whole data set. As shown in Figure 3, the fit statistic INFIT MNSQ (Linacre & Wright, 1994) are very close to its expected value, 1.0.

Item 2 has the largest fit statistic. It may call for further investigation. Generally speaking, all the ten items fit the partial credit model fairly well. It should be noted that model-data fit is a matter of degrees rather than all or none. To check if these items show Gender main effects, Age main effects, or Gender by Age interaction effects, a full model and several reduced models were conducted. In the full model, all items are assumed to show all kinds of DIF effects on all steps, which leads to 57 DIF parameters, in addition to three mean-deviation parameters, 19 grand step difficulties, and two person distribution parameters. This model has a likelihood statistic (= -2 × loglikelihood) of 33045.61, with 81 parameters. The estimated parameters of this model are listed as the generating values in Table 1. Next, several reduced models were formed by constraining one of the DIF parameters to zero consecutively. The likelihood ratio test was applied to compare the full model with the reduced models. The form of the likelihood ratio test is

$$G_{df}^2 = 2\left(\text{loglikelihood (F)} - \text{loglikelihood (R)}\right),$$

where loglikelihood $(\cdot)$ represents loglikelihood of the data given the maximum likelihood estimates of the parameters of the model; *df* is the difference between the number of parameters in the full model and that of the reduced model. G_{df}^2 follows approximately the chi-squared distribution with *df* degrees of freedom when the reduced model is true (Rao, 1973).

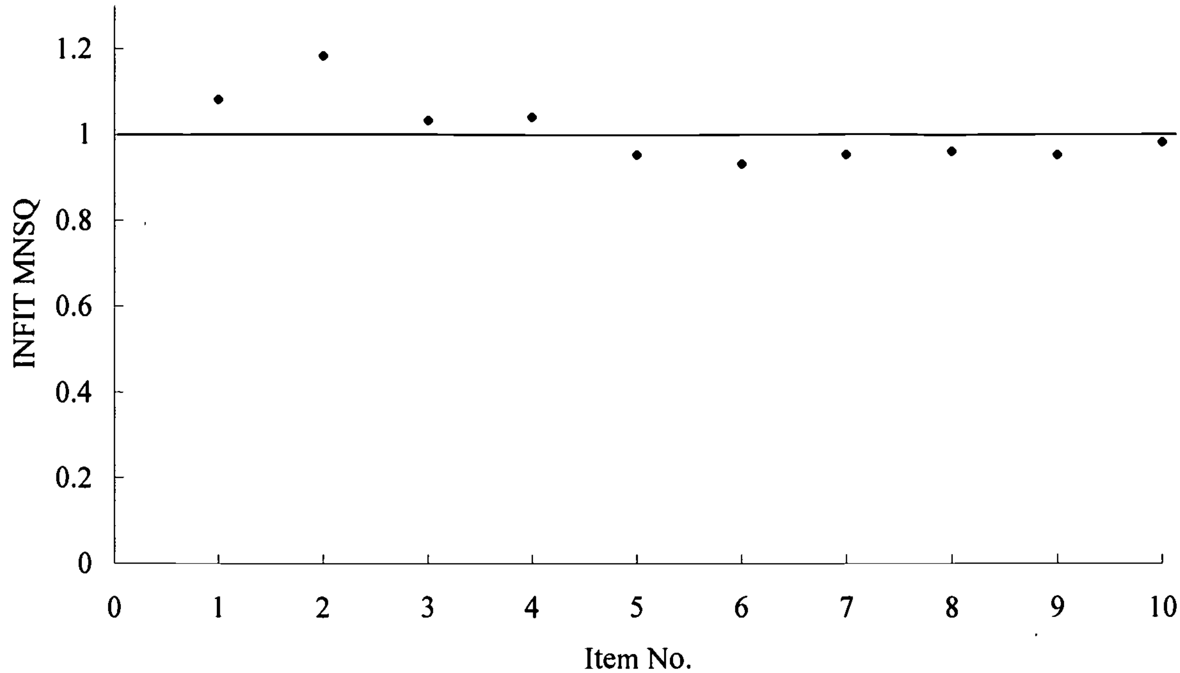

Figure 3. Fit statistics of the ten polytomous items to the partial credit model

In this case, *df* is equal to 1. If G_1^2 is greater than the critical value at the .05 level, 3.84, the step of the item shows a particular kind of DIF. According to Table 3, of the 57 DIF parameters, 22 parameters are significantly different from zero at the .05 level. Another reduced model where these 22 DIF parameters were estimated and the other DIF parameters were constrained to zero was then conducted. This model has a likelihood statistic of 33093.42, with 46 parameters. According to the likelihood ratio test, this reduced model is not significantly different from the full model ($G_{35}^2 = 47.81, p = .07$) and thus is preferred. In sum, out of the 57 possible DIF parameters for the ten polytomous items across the four groups, 22 parameters are statistically different from zero.

The estimated parameters of this reduced model are listed as the generating values in Table 2. Consider the means of the four groups. The first mean-deviation (Gender main effect) and the third mean-deviation (Gender by Age interaction effect) are trivial, -.04 and -.02, respectively. The second deviation (Age main effect) is relatively large, .18. Accordingly, the major difference among the four groups is between the young and the old, .36 (= .18 × 2). The means of the four groups can be obtained with Equation (10):

Young Males: .37 + (-.04) + .18 + (-.02) = .49

Young Females: .37 - (-.04) + .18 - (-.02) = .61

Old Males: .37 + (-.04) - .18 - (-.02) = .17

Old Females: .37 - (-.04) - .18 + (-.02) = .21

Consequently, the young females are the most liberal (i.e., putting less values on family) and the old males are the most conservative (i.e., putting high values on family).

Table 3. Likelihood ratio tests for various DIF parameters

Item_step	G_1^2	Item_step	G_1^2
Main effects of factor A_1 (Males)		Interaction effect of factors A_1 (Males) and B_1 (Young)	
1_1	1.34	1_1	1.55
1_2	5.45*	1_2	6.39*
2_1	1.13*	2_1	.67
2_2	28.39*	2_2	7.62*
3_1	5.31*	3_1	1.02
3_2	7.49*	3_2	.88
4_1	.21	4_1	1.81
4_2	6.82*	4_2	.42
5_1	.08	5_1	3.82
5_2	2.86	5_2	.7
6_1	.01	6_1	2.82
6_2	.1	6_2	1.04
7_1	0	7_1	4.15*
7_2	5.45*	7_2	5.25*
8_1	.07	8_1	7.20*
8_2	3.51	8_2	0
9_1	3.4	9_1	.2
9_2	1.2	9_2	3.82
10_1	2.04	10_1	.66
10_2	3.56	10_2	3.81
Main effects of factor B_1(Young)			
1_1	9.27*		
1_2	5.53*		
2_1	2.64		
2_2	17.16*		
3_1	.19		
3_2	11.22*		
4_1	4.70*		
4_2	32.30*		
5_1	2.55		
5_2	23.09*		
6_1	.19		
6_2	.23		
7_1	1.29		
7_2	4.93*		
8_1	.35		
8_2	8.22*		
9_1	8.02*		
9_2	2.45		
10_1	.13		
10_2	3.74		

* $p < .05$

The delta scale can be converted into the logit scale as follows. As stated, a difference of 1.0 delta corresponds to a difference of 10 points in percentage correct between groups. Assume the two groups have percentages correct of .45 and .55, respectively. Also assume the ability levels for the two groups are both 0.0 logits. According to the Rasch model, it leads to:

$$\log\left(\frac{.45}{.55}\right)_1 = 0 - \delta_{i1},$$

$$\log\left(\frac{.45}{.55}\right)_2 = 0 - \delta_{i2},$$

where subscripts 1 and 2 denote the two groups. Consequently, δ_{i1} and δ_{i2} are .20 and -.20 logits, respectively. The difference of the two difficulties is .40 logits. Therefore, a difference of 1.0 delta corresponds roughly to .40 logits. Likewise, 1.5 deltas corresponds roughly .60 logits. Therefore, if the difference of two item difficulties between groups is smaller than.40 logits, this item is in category *A*. If the difference is larger than .60 logits, it goes to category *C*. All other items belong to category *B*.

Draba (1977) provided another rule of classification: An item is identified as exhibiting substantial DIF if the difference of item difficulty estimates for any two groups was more than .50 logits. Obviously, these two rules of classification are quite similar. Although these two rules were derived for dichotomous items, they might be applied to polytomous items, because the step difficulties in the partial credit model are directly extended from the item difficulties in the Rasch model. Given no rules for polytomous items are available in the literature and any rule in some sense is arbitrary, Draba's is used in this paper. Since there is more than one step for polytomous items, if any step of an item exhibits substantial DIF on any two groups, the item is said to have DIF.

Given there are only two levels in each factor, the DIF step parameters are directly related to the differences. Among the significant DIF parameters, if one is less than .25 (= .50 /2) logits, the corresponding step exhibits substantial DIF. Of the 22 significant DIF parameters, as shown in Table 3, only five steps exhibit substantial DIF, which come from items 2, 4, and 5. Consider item 2 as an example:

If my siblings ask me to be their financial guarantor, I should never reject them.

The main effects of Gender on the first and the second step are .43 and .34, respectively. In other words, given identical levels on the trait, at the first step (from "strongly agree" to "agree"), the item is .86 (= .43 × 2) logits more difficult for the males than for the females, given identical levels on the trait. At the second step (from "agree" to "not agree"), the item is .68 (= .34 × 2) logits more difficult for the males than for the females, given identical levels of the trait. That is, the probability of choosing "agree" rather than "strongly agree" (scoring 1 rather than 0), and that of choosing "not agree" rather than "agree" (scoring 2 rather than 1) are both lower for the males than for the females with identical levels on the trait. In Chinese society, females usually do not hold close relationship with their siblings once they get married. They usually have little power over home finance. However, the relationship with siblings for males does not change remarkably when they get married. Therefore, adult females and males may have quite different perspectives on serving financial guarantors for their siblings. This may partly account for the main effects of Gender.

The main effect of Age on the second step is .27. It means that at the second step, the item is .54 (= .27 × 2) logits more difficult for the young than for the old. That is, the probability of choosing "not agree" rather than "agree" is lower for the young than for the old with identical levels on the trait. The other DIF parameters can be interpreted in the same way.

The first and the second step difficulties of item 2 for the four groups can be obtained with Equation (8) as follows:

The first step:

Young Males: -2.28 + .43 + 0 + 0 = -1.85

Young Females: -2.28 - .43 + 0 - 0 = -2.71

Old Males: -2.28 + .43 - 0 - 0 = -1.85

Old Females: -2.28 - .43 - 0 + 0 = -2.71

The second step:

Young Males: -1.10 + .34 + .27 + .20 = -.29

Young Females: -1.10 - .34 + .27 - .20 = -1.37

Old Males: -1.10 + .34 - .27 - .20 = -1.23

Old Females: -1.10 - .34 - .27 + .20 = -1.51

Figure 4 shows the expected scores on item 2 for the four groups. The males and the females have quite different curves (i.e., expected scores). The young females and the old females have almost identical curves. The young males and the old males have somewhat different expected scores, which results from a substantial main effect of Age and a marginal interaction effect (.20) on the second step. If all the DIF parameters in an item are zero, the expected scores on that item for the four groups will be identical.

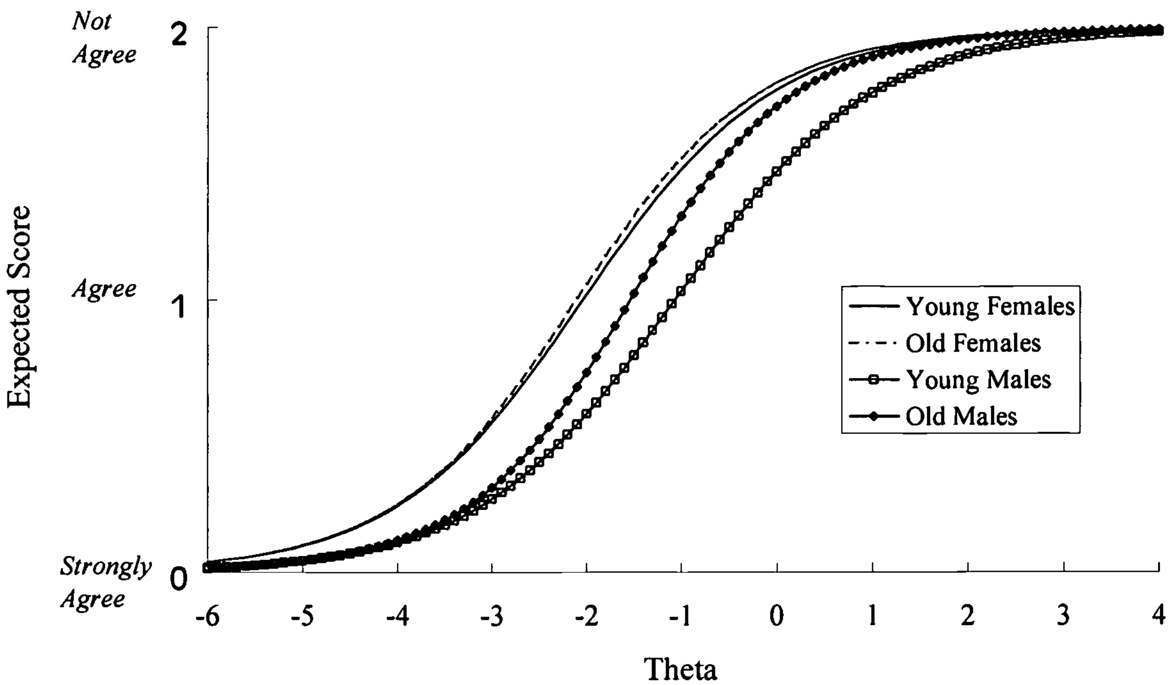

Figure 4. Expected score on item 2 for the four groups

Conclusion

In analysis of DIF with more than one grouping factor, we may either do marginal DIF analysis by collapsing across the groups, or treat all the possible group combinations as drawn from a unified factor. In doing so, interaction effects among the factors become invisible or the definition of the unified factor is not well defined so that the interpretation of DIF is vague. In this study, a procedure that jointly analyzes all groups while holding individual factors is proposed. It is based on the formulation of general linear models. Item parameters across all groups are reparameterized as a set of grand item parameters and several sets of parameters representing main and interaction effects of the factors on items. With this parameterization, test users are able to investigate thoroughly how items are affected by the factors and how they interact. This information can help revise those items with substantial DIF and clarify the constructs that underlie subjects' responses.

Simulation studies were conducted under two conditions: a full model where all possible DIF parameters were estimated and a reduced model where only parts of DIF parameters were estimated. Results show that all the parameters were recovered very well and no systematic patterns of bias were found, which suggests that the proposed procedure is not only theoretical preferable but also applicable. A real data set of ten polytomous items and 1924 subjects was analyzed. Two factors were formed: Gender and Age. No Gender by Age interaction effects on any step were substantial. Item 2 has the main effect of Gender on two steps. Items 2, 4 and 5 have the main effects of Age on their second steps. With this information, test developers or users are able to investigate DIF effects and revise item when needed. Although in this study, two factors with two groups in each are illustrated, this approach can be directly generalized to more than two factors with more than two groups in each.

References

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443-459.

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A programmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135-165). Hillsdale, NJ: Erlbaum.

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 26). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59,* 177-192.

Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling: Handbook of the Usage of LPCM-Win 1.0.* Groningen, the Netherlands: ProGAMMA bv.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Linacre, J. M., & Wright, B. D. (1994). *A user's guide to BIGSTEPS: Rasch-model computer program.* Chicago: MESA Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.*

Hillsdale, NJ: Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1993). *Implementing item parameter drift and bias in polytomous item response models.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta GA.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligent and attainment tests.* Copenhagen: Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Rogers, H. J., & Swaminathan, H. (1993). *Differential item functioning procedures for non-dichotomous responses.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta GA.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braum (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Wang, W.-C., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard, & K. Draney (Eds.),

TM029709

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Decomposition of DIF Effects for Polytomous Items	
Author(s): Wen-Chung WANG	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixe Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTE SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
✓	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitti reproduction and dissemination in microfiche

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.	
Signature: [signature]	Printed Name/Position/Title: WEN-CHUNG WANG/Associate Prof.
Organization/Address: Dept. of Psychology. Chung Cheng University, Chia-Yi 621 Taiwan	Telephone: 886-5-2720411 #6430 / Fax: 886-5-2720857
	E-mail Address: psywcw@ccunix.ccu.edu.tw / Date: 04/15/99

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:	
ERIC Clearinghouse on Assessment and Evaluation 1129 Shriver Laboratory (Bldg 075) College Park, Maryland 20742	**Telephone: 301-405-7449 Toll Free: 800-464-3742 Fax: 301-405-8134 ericae@ericae.net http://ericae.net**

EFF-088 (Rev. 9/97)